
Exercise 1: PCA and LDA (30 points). Consider 2D data points coming from a mixture of two Gaussians with equal proportions, different means, and equal, diagonal covariances (where $\mu, \sigma_1, \sigma_2 > 0$):

$$ x \in \mathbb{R}^2: p(x) = \pi_1 p(x|1) + \pi_2 p(x|2) \quad p(x|1) \sim \mathcal{N}(\mu_1, \Sigma_1), \quad p(x|2) \sim \mathcal{N}(\mu_2, \Sigma_2), $$

$$ \pi_1 = \pi_2 = \frac{1}{2}, \quad \mu_1 = 0, \quad \mu_2 = \left( \begin{array}{c} \mu \\ 0 \end{array} \right), \quad \Sigma_1 = \Sigma_2 = \left( \begin{array}{cc} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{array} \right). $$

1. (5 points) Compute the mean $\mu$ and covariance $\Sigma$ of $p$. Hint: see exercise below.
2. (5 points) Compute the eigenvalues $\lambda_1 \geq \lambda_2 \geq 0$ and corresponding eigenvectors $u_1, u_2 \in \mathbb{R}^2$ of $\Sigma$. Can we have $\lambda_2 > 0$?
3. (2 points) Find the PCA projection to dimension 1.
4. (5 points) Compute the within-class and between-class scatter matrices $S_W, S_B$ of $p$.
5. (6 points) Compute the eigenvalues $\nu_1 \geq \nu_2 \geq 0$ and corresponding eigenvectors $v_1, v_2 \in \mathbb{R}^2$ of $S_W^{-1}S_B$. Can we have $\nu_2 > 0$?
6. (2 points) Compute the LDA projection.
7. (5 points) When does PCA find the same projection as LDA? Give a condition and explain it.

Exercise 2: mixture distributions (10 points). Let $p(x) = \sum_{k=1}^{K} \pi_k p(x|k)$ for $x \in \mathbb{R}^D$ be a mixture of $K$ densities, where $\pi_1, \ldots, \pi_K \in [0,1]$ and $\sum_{k=1}^{K} \pi_k = 1$ are the component proportions (prior probabilities) and $p(x|k)$, for $k = 1, \ldots, K$, the component densities (e.g. Gaussian, but not necessarily). Let $\mu_k = E_{p(x|k)} \{ x \}$ and $\Sigma_k = E_{p(x|k)} \{ (x - \mu_k)(x - \mu_k)^T \}$ be the mean and covariance of component density $k$, for $k = 1, \ldots, K$.

1. (5 points) Prove that the mean and covariance of the mixture are:

$$ \mu = E_{p(x)} \{ x \} = \sum_{k=1}^{K} \pi_k \mu_k \quad \Sigma = E_{p(x)} \{ (x - \mu)(x - \mu)^T \} = \sum_{k=1}^{K} \pi_k \left( \Sigma_k + \mu_k \mu_k^T \right) - \mu \mu^T. $$

2. (5 points) Imagine the component covariances $\Sigma_1, \ldots, \Sigma_K$ are all diagonal. Is the mixture covariance diagonal? Explain.
Exercise 3: variations of k-means clustering (30 points). Consider the k-means error function:

\[ E(\{\mu_k\}_{k=1}^K, Z) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \| x_n - \mu_k \|^2 \quad \text{s.t.} \quad Z \in \{0, 1\}^{NK}, \quad \mathbf{1} = 1 \]

over the centroids \( \mu_1, \ldots, \mu_K \) and cluster assignments \( Z_{N \times K} \), given training points \( x_1, \ldots, x_N \in \mathbb{R}^D \).

- **Variation 1**: in k-means, the centroids can take any value in \( \mathbb{R}^D \): \( \mu_k \in \mathbb{R}^D \) \( \forall k = 1, \ldots, K \). Now we want the centroids to take values from among the training points only: \( \mu_k \in \{x_1, \ldots, x_N\} \) \( \forall k = 1, \ldots, K \).
  1. (8 points) Design a clustering algorithm that minimizes the k-means error function but respecting the above constraint. Your algorithm should converge to a local optimum of the error function. Give the steps of the algorithm explicitly.
  2. (2 points) Can you imagine when this algorithm would be useful, or preferable to k-means?

- **Variation 2**: in k-means, we seek \( K \) clusters, each characterized by a centroid \( \mu_k \). Imagine we seek instead \( K \) lines (or hyperplanes, in general), each characterized by a weight vector \( w_k \in \mathbb{R}^D \) and bias \( w_{k0} \in \mathbb{R} \), given a supervised dataset \( \{(x_n, y_n)\}_{n=1}^N \) (see figure). Data points assigned to line \( k \) should have minimum least-squares error \( \sum_{n \in \text{line } k} (y_n - w_k^T x_n - w_{k0})^2 \).
  1. (8 points) Give an error function that allows us to learn the lines’ parameters \( \{w_k, w_{k0}\}_{k=1}^K \).
  2. (12 points) Give an iterative algorithm that minimizes that error function.

Exercise 4: mean-shift algorithm (10 points). Consider a Gaussian kernel density estimate

\[ p(x) = \sum_{n=1}^N p(x|n)p(n) = \frac{1}{N(2\pi\sigma^2)^{D/2}} \sum_{n=1}^N e^{-\frac{1}{2} \| \frac{x - x_n}{\sigma} \|^2} \quad x \in \mathbb{R}^D. \]

Derive the mean-shift algorithm, which iterates the following expression:

\[ x \leftarrow \sum_{n=1}^N p(n|x)x_n \quad \text{where} \quad p(n|x) = \frac{p(x|n)p(n)}{p(x)} = \frac{\exp \left( -\frac{1}{2} \| (x - x_n)/\sigma \|^2 \right)}{\sum_{n'=1}^N \exp \left( -\frac{1}{2} \| (x - x_{n'}/\sigma \|^2 \right)} \]

until convergence to a maximum of \( p \) (or, in general, a stationary point of \( p \), satisfying \( \nabla p(x) = 0 \)).

**Hint**: take the gradient of \( p \) wrt \( x \), equate it to zero and rearrange the resulting expression.

Exercise 5: nonparametric regression (20 points). Consider the Gaussian kernel smoother

\[ g(x) = \sum_{n=1}^N \frac{K\left(\| (x - x_n)/h \| \right)}{\sum_{n'=1}^N K\left(\| (x - x_{n'}/h \| \right)} y_n \quad \text{where} \quad K\left(\| x - x_n/\sigma \| \right) \propto \exp \left( -\frac{1}{2} \| (x - x_n)/\sigma \|^2 \right) \]

estimated on a training set \( \{(x_n, y_n)\}_{n=1}^N \subset \mathbb{R}^D_x \times \mathbb{R}^D_y. \)

1. (7 points) What is \( g(x) \) if the training set has only one point \((N = 1)\)? Explain. Sketch the solution in 1D (i.e., when both \( x_n, y_n \in \mathbb{R} \)). Compare with using a least-squares linear regression.

2. (13 points) Prove that, with \( N = 2 \) points, we can write \( g(x) = \alpha(x)y_1 + (1 - \alpha(x))y_2 \) where \( \alpha(x) \) can be written using the logistic function. Give the detailed expression for \( \alpha(x) \). Sketch the solution in 1D. Compare with using a least-squares linear regression.